心理科学进展 2019, Vol. 27, No. 3, 475–489
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2019.00475

·研究前沿(Regular Articles)·

语言加工过程中的视听跨通道整合*

韩海宾 许萍萍 屈青青 程 茜 李兴珊
(中国科学院心理研究所, 北京 100101) (中国科学院大学心理学系, 北京 100049)

摘 要 日常生活中, 语言的使用往往出现在某个视觉情境里。大量认知科学研究表明, 视觉信息与语言信息加工模块并不是独立工作, 而是存在复杂的交互作用。本文以视觉信息对语言加工的影响为主线, 首先对视觉信息影响言语理解, 言语产生以及言语交流的相关研究进展进行了综述。其次, 重点对视觉信息影响语言加工的机制进行了探讨。最后介绍了关于视觉信息影响语言加工的计算模型, 并对未来的研究方向提出了展望。

关键词 视觉信息; 语言加工; 言语理解; 言语产生; 言语交流
分类号 B842

1 引言

日常生活中, 人们经常同时接受来自不同感觉通道的信息。例如, 当人和人面对面交流时, 人们的耳朵在听到话语的同时, 眼睛能同时看到相关的视觉信息。人们在加工这些来自不同通道的信息时, 往往利用不同的认知模块。近代认知神经科学的研究表明, 人脑也往往利用不同的脑区对不同通道的信息进行加工(Binder et al., 1997; Grill-Spector & Malach, 2004)。然而, 也有研究发现不同的认知模块往往不是在独立工作, 而是相互影响的(Beauchamp, 2016; Kuchenbuch, Paraskevopoulos, Herholz, & Pantev, 2014; Marslen-Wilson, 1975; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Eggermont, 2017)。举例来说, 语言的“意有所指”众所周知, 我们听到的口语词汇往往对应着视觉世界中的特定物体。因此, 在同时加工口语和视觉信息时, 语言会引导视觉注意, 视觉信息也会影响语言加工, 听觉与视觉通道的信息相互影响, 共同完成整合任务。近年来, 随着计算机技术的不断发展, 人工智能也成为了研究热点。研究者也开始尝试把不同通道的信息整合进人工智能, 使其完成更为复杂的功能, 更好地为人类服务(Ng et al., 2017; Heinrich & Wermter, 2017)。

加工来自不同感觉通道信息的认知模块如何相互影响, 又如何完成来自不同通道的信息整合任务, 是认知心理学需要研究的重要问题。而针对多通道信息整合原因以及整合机制的问题, 目前研究尚浅。本文将重点综述近年来针对视觉信息的加工如何影响口语信息加工的研究进展。首先通过介绍语言加工的模块化理论以及交互理论两大理论来引出争议问题; 其次介绍视觉信息影响语言加工的表现以及为何会影响语言加工两个大问题; 最后将介绍视觉影响语言加工的计算模型, 并对未来的研究进行展望。

2 模块化理论与基于制约的理论

语言的加工包含语言理解、语言产生等多种加工过程。语言理解、言语产生过程又细分为词汇识别、句法解析、言语计划等过程。这些语言加工的过程是独立进行, 还是会受到其他信息的影响, 长久以来都是心理语言学家们互相争论的热点。上世纪 80 年代早期, Fodor (1983)提出了语言的“模块化理论”, 该理论认为人脑的认知系统由许多不同的模块构成。例如, 在语言加工系统

收稿日期: 2018-02-28
* 国家自然科学基金委与德国科研基金委联合资助项目(NSFC 61621136008/DFC TRR-169)及国家自然科学基金委项目(31571125, 31771212)资助。
通信作者: 李兴珊, E-mail: lixs@psych.ac.cn

中，有负责语音加工的模块，有负责词汇加工的模块，有负责句法加工的模块等等。每个模块都是独立的加工单位，其活动与输出不受其他信息的影响。举例来讲，句子的理解过程包含对句子语义信息的通达、句法结构的建构等过程。根据模块化理论观点，句法加工模块独立于语境、语义等信息的加工模块，负责句法加工的模块是“封装”起来的，不受其他高级认知或者感知觉机制的影响。当然，模块化理论并非不承认高水平的信息(例如语境)对句法加工的影响，在遇到一词多义或者歧义现象时，同样需要根据语境等信息来确定歧义词在句子中的语义。其倡导的主要观点是：高水平信息无法影响句法加工的最初阶段，但是会在句法加工的最初阶段完成后给予反馈，而不是直接参与到句法初级阶段的加工过程中来。

支持模块化理论的代表模型是花园路径模型。该模型由 Frazier 和 Rayner (1982)提出，并得到其实验结果的支持。作者认为对任何一个歧义句的加工最开始只考虑一种可能的句法结构，并且最初句法结构的选择只是纯粹的句法加工模块起作用，之后出现加工困难之后才会依据语境、语义等信息进行反馈。但 Altmann, Garnham 和 Dennis (1992)通过严格地控制语境，使语境符合歧义句多种句法结构的其中一种，发现恰当的语境可以移除歧义句中的加工困难，直接选择与语境相符的句法结构。Altmann 等人(1992)的结果并不支持模块化理论，作者认为句法加工模块并非无法“渗透”，也并没有“封装”起来，语境这种高水平的信息可以自上而下地影响句法的最初选择策略。

除此之外，模块化理论出现之前已有研究发现语言加工系统的加工器之间可以互相传递信息，各种加工过程之间也会相互影响。例如, Marslen-Wilson (1975)发现句法水平和语义水平的信息可以影响词汇识别过程。作者采用影子跟读任务(the shadowing task)考察了语境对词汇识别以及词汇整合过程的影响。被试需要听句子并及时对听到的词汇进行复述(跟读任务)。目标词的类型分为语义反常、句法反常以及正常词汇三个条件，例如目标词“universe (宇宙)”在句子“the new peace terms have been announced. They call for the unconditional universe of all the enemy force (新的和平条款已经宣布了。他们呼吁所有的敌人无条件的宇宙”中属语义反常的词汇; “already (已经)”在句子“he thinks she won't get the letter. He's afraid he forgot to put a stamp on the already before he went to post it (他认为她不会收到那封信。恐怕他在去邮寄之前忘了在已经上贴邮票)”中属句法反常词汇。每种条件下的词汇又分为四组，一组为原始词汇(universe)，其余三组分别为替换掉首音节(u)、次音节(n)以及三音节(i)的非词。结果发现和语境相匹配的无反常条件下，次音节与三音节替换组的跟读错误率最高，即，在首音节不变而且符合语境的条件下被试会将其跟读成正常词，说明语境这种高水平的信息确实会影响词汇的识别过程，与之发生交互作用。

以上研究都支持在语言加工过程中，句法和语义是有交互作用的。其中一种交互作用的观点被称为约束满足理论(constraint satisfaction theory)或者基于制约的模型(constraint-based model) (MacDonald, 1993; MacDonald, Pearlmutter, & Seidenberg, 1994)。该模型强调了语言加工中各类信息即时相互作用，认为语境、句法使用频率等信息可以即时被句法加工所使用，初级阶段的句法选择也会受到影响，整个句子的建构过程是各种信息交互作用、相互制约的结果。例如，在歧义句理解中，可供选择的句法是平行的，会受到语境信息、句法使用频率以及语义等信息的制约。歧义消解则是一个约束满足的过程，语境、句法使用频率等信息提供证据支持部分被激活的句法结构。该模型得到了众多研究的支持(Chen & Tsai, 2015; Knoeferle & Guerra, 2016; Linzen & Jaeger, 2016; MacDonald, 1993)。除语境、频率等语言类信息之外，还有突显性更高的视觉情境等非语言信息也可能会影响句法加工过程，但由于实验技术等方面的原因，早期对这个问题的考察较少。视觉情境范式的广泛应用使得这类研究如雨后春笋般涌现出来。

3 视觉信息影响语言加工的表现

已有很多来自视觉情境范式的研究证据支持语言的加工会受到视觉场景等非语言信息的影响。视觉情境范式(the visual world paradigm, VWP)的出现为考察视知觉信息与词汇、句法和语义等更高级语言加工的交互作用打开了一扇大门。这种范式最突出的特点就是在被试观看视觉刺激的

同时向被试呈现听觉语言信息，要求被试根据听到内容选择相对应的视觉刺激或对物体进行一定的操作，或者单纯地听和看。通过记录被试的眼动来评估语言加工过程中视觉注意的分配情况，进而对语言加工的机制进行推论。视觉情境范式由 Roger M. Cooper 于 1974 年首创，他向被试呈现一些物体图片，同时播放一些短文录音，发现在听到某个特定的词语时，被试会更多地注视与听觉信息具有语义关系的图片。比如，在听到"非洲(Africa)"时，对语义上相关的"斑马(zebra)"、"狮子(lion)"以及"蛇(snake)"等物体的注视比例比无关的物体更多，而且被试的眼动与文本的听觉呈现在时间上是紧密相关的。通过视觉情境范式我们不仅可以揭示语言的加工机制，而且能够考察视觉信息如何影响语言的加工过程。

这部分主要介绍关于视觉信息影响语言加工的一些经典研究，主要从视觉信息影响口语理解以及言语产生两个方面来综述视觉信息影响语言加工的表现。除此之外，视觉情境范式里不仅有听觉呈现的语言，而且还有视觉画面的呈现，这和单纯的语言加工过程也存在差异，这里将会对此类现象进行一些探讨。

3.1　视觉信息影响口语加工过程

视觉信息影响音节层面的口语信息加工。早期发现的"麦格克效应" (the McGurk Effect)就已经发现了音节层面视觉与听觉之间的交互作用(McGurk & MacDonald, 1976)。实验任务要求被试看到一个面孔重复发音"ga"的嘴部动作，并听到和视频中嘴部动作同时出现的声音"ba"。结果发现，虽然听觉输入可以非常清楚地被知觉为"ba"，但由于视觉输入的面孔口型的影响会让被试知觉为"da"，表明视知觉对音节的感知有干扰作用。

视觉信息可以影响单个词汇的理解过程。Tanenhaus 等(1995)首次使用视觉情境范式对视觉信息如何影响单个词汇上的暂时歧义消解进行了探究。实验过程中，被试需要在听到词汇"candy (糖果)"的同时看包含一些物体的图片，这些图片分为两组：一组包含目标物"candy"以及干扰项；另一组包含目标物"candy"，与"candy"具有相同起始音节的竞争物"candle (蜡烛)"以及干扰项。作者发现如果视觉画面中不呈现竞争物"candle"，被试指向目标物"candy"的眼跳潜伏期为 145 ms，如果同时呈现目标物"candy"和竞争物"candle"，指向目标物的眼跳潜伏期则变为 230 ms，显著长于不呈现竞争物的条件。作者认为被试听到"can-"的时候会产生临时歧义，视觉信息中竞争物的呈现影响了被试在词汇水平上的临时歧义消解过程。在词汇水平上，Chambers, Tanenhaus, Eberhard, Carlson 和 Filip (1998)发现视觉场景信息中的语用信息(pragmatic factor)和介词语义信息会共同作用来缩小介宾短语中宾语的指涉范围。在实验中，作者给被试呈现包含有可以装下"cube"的"big can (大容器)"，不可以装下"cube"的"small can (小容器)"以及其他干扰物的视觉场景，同时让被试听句子"put the cube inside the can (把方块放到容器中)"。结果发现被试并不是注视所有容器类物体，而是直接注视场景中可以装下"cube"的大容器。这是因为视觉场景含有"哪个容器可以放下方块"的语用信息，而且这些语用信息影响了个体对介词"inside (里面)"的理解，缩小了介宾短语中宾语的指涉范围。

除影响单个词汇的理解过程之外，视觉信息还会影响句法加工过程，这也是使用视觉情境范式考察最多的一部分。Tanenhaus 等人(1995)的研究开创了先河，对视觉信息如何影响歧义句中的句法选择进行了考察，并为视觉信息对句法加工的影响提供了充足的证据。该研究选取局部歧义句作为听觉实验材料，例如，句子"Put the apple on the towel in the box (把毛巾上的苹果放到盒子里)"中临时歧义部分是短语"on the towel (在毛巾上)"，既可以修饰名词"apple (苹果)"，意为"毛巾上的苹果"，也可以指向"put (放)"的目标位置，意为"把苹果放到毛巾上"。和听觉刺激同时呈现的视觉刺激包含两种条件(如图 1)，单表征物情境(1-referent, 左图)以及双表征物情境(2-referents, 右图)。作者的假设为不同的视觉情境会使被试有不同的句法选择策略，即两种视觉刺激条件下会对歧义短语有不同的理解，并表现为不同的眼动模式。具体来说，单表征物情境条件下，由于只有一个目标物，所以被试更倾向于将"on the towel"理解成和动词"put"相关的目的地，会有更多的错误注视在毛巾上；在双表征物情境条件下，由于存在两个目标物，被试需要选择其中一个做为接受动作的客体，会更多地将"on the towel"加工成"apple"的修饰语，表现为对毛巾错误注视概率的减少。结果验证了其假设，在句法加工早期单表

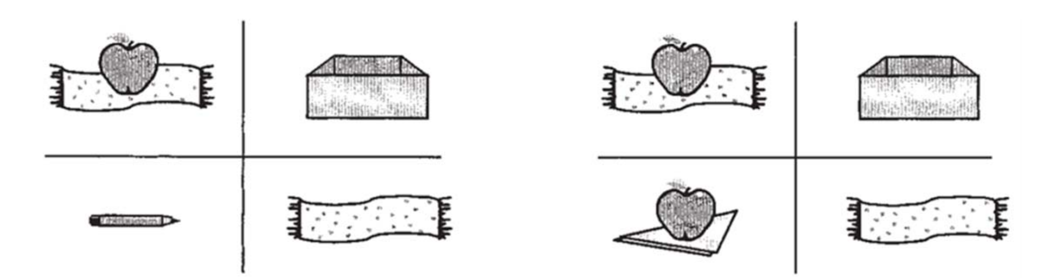


图 1 Tanenhaus 等(1995)使用的视觉刺激。左图为单表征物情境，右图为双表征物情境，被试在看图片的同时会听到局部歧义句“Put the apple on the towel in the box”。

征物情境条件下的错误注视概率要显著多于后者，视觉情境参与了句法加工的早期阶段，并且移除了临时歧义句中的加工困难。作者认为句法加工过程并非如模块化理论所倡导的不受到其他信息影响，视觉情境提供的信息可以即时地影响大脑对句法结构的选择策略，并即时地应用在句子结构歧义的消解上。这与 Altmann 等人(1992)的研究结果是一致的，从“语境”变为了“视觉情境”，都支持了基于制约的理论。

很多研究利用相似的范式对这个问题进行了重复与拓展。一些研究采用和其相同的电脑屏幕呈现的图片形式，还有部分研究采用呈现实物的方法来代替电脑呈现。例如，有研究考察了儿童与成人在视觉与语言加工整合上的差异，结果发现成人可以将语言信息(如，词汇信息)和表征物信息(视觉信息)有效地结合，移除句子中的临时歧义；儿童却只能利用听觉句子中的语义和句法信息来进行句子理解，对视觉信息的利用是十分有限的(Snedeker & Trueswell, 2004)。

除视觉信息之外，其他非语言信息如物体的动允性、事件以及情景记忆都会和语言的加工发生交互作用(Chambers & Juan, 2008; Chambers, Tanenhaus, & Magnuson, 2004; Lee, Chambers, Huettig, & Ganea, 2017; Leonard & Chang, 2014; Milburn, Warren, & Dickey, 2015)。例如, Chambers 等人(2004)采用视觉情境范式考察了非语言信息(动允性，affordance，指的是环境的属性使得动物个体的某种行为得以实施的可能性，Eysenck & Keane, 2000)对局部歧义句句法加工过程的影响。实验中，作者给被试听指导语“Pour the egg in the bowl over the flour (把碗里的鸡蛋放到面粉上)”，其中，“in the bowl (在碗里)”既可以修饰名词“egg (鸡蛋)”，意为“碗里的鸡蛋”，也可以指向“pour (倒)”的目标位置，意为“把鸡蛋倒入碗里”。同时呈现两种真实的视觉场景并让读者根据听到的指导语操作物体(如图 2)。作者构建了两个条件：一个是竞争物和目标物都是液体形式的鸡蛋(都可以被倒在面粉上，具有“pour (倒)”的动允性)；另一条件下只有一个鸡蛋是液体形式。结果表明，在第二个条件的场景下被试对“bowl (碗)”的错误注视概率会更高，更容易把“in the bowl (在碗里)”理解成行为的目的，这表明和动作相关的非语言信息影响了句法的早期加工过程。

情节记忆同样会影响语言的加工过程(Chambers

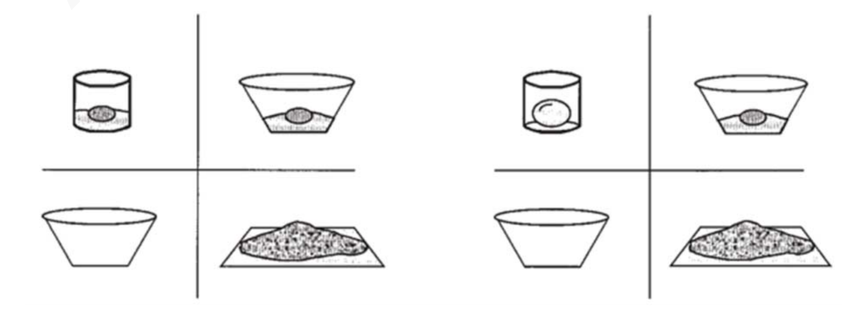


图 2 Chambers 等(2004)使用的视觉刺激示例。左图为包含两个液体鸡蛋的双表征物情境，右图为只包含一个液体鸡蛋的单表征物情境。

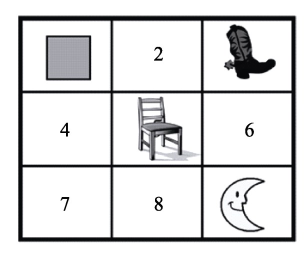


图 3　Chambers 和 Juan (2008)的研究使用的视觉刺激示例。从方形到月亮位置标号依次为 1~9。

& Juan, 2008; van Bergen & Flecken, 2017)。例如，在 Chambers 和 Juan (2008)的研究中，被试需要看如图 3 所示图片的同时听三个指导语，分别为“move the chair to area two (把椅子移到区域 2)”, “now move/return the chair to area five (把椅子移到/放回区域 5)”, “now move the square to area seven (把方块移到区域 7)”。其中第二个指导语是关键指导语，分为“move (移动)”和“return (放回)”两个实验条件，其中“return”条件需要第一个指导语产生的情节记忆的参与。结果发现，在“return”条件下，被试在听到“return”的时候会出现指向椅子未移动之前的区域 5 的预期眼跳，“move”条件下则没有。实验表明听者的预期不仅仅基于物体的特点，视觉场景产生的情节记忆也同样会影响被试对句子的加工过程。这些都为非语言信息影响句子加工过程提供了进一步的证据。

此外，不仅静态视觉信息对句法加工有影响，动态的事件也会影响口语理解过程(Hafri, Trueswell, & Strickland, 2018; Knoeferle & Guerra, 2016; Knoeferle, Crocker, Scheepers, & Pickering, 2005)。Knoeferle 等人(2005)采用视觉−情境范式考察了图片所呈现事件情境是否可以影响口语句子中题元角色的分配(thematic-role assignment)，即是否影响被试在句子加工中施动者(agent)和受动者(patient)的角色分配。在实验过程中，给被试呈现一个视觉事件，如图 4 所示，“princess (公主)”处在一个既在给“pirate (海盗)”清洗，同时又被“fencer (击剑者)”所画的两种角色状态下(即公主既可能是受动者也可能是施动者)。同时以听觉形式给被试呈现两种条件的指导语“the princess is apparently washing the pirate (公主很明显在清洗海盗)”和“the princess is apparently painted by the fencer (公主很明显在被击剑者画)，前者公主作为施动者，后者为受动者。结果发现，前者条件下，被试听到动词“washing (洗)”会出现更多的指向海盗的预期眼动，后者则更多的看向击剑者。实验表明被试已经从视觉情境中提取出该事件的题元角色的分配情况，一旦动词出现，句子的题元角色分配就已经完成。作者认为视觉画面中提取的题元角色信息加速了口语理解中题元角色的分配，描述某事件的视觉场景促进了口语理解的过程。

综上所述，不仅静态图片和真实情境能够影响我们对听觉语言信息的加工，动态事件情境信息也同样会影响语言的理解过程。这种影响不仅体现在单个词汇水平，同样表现在语言加工过程中的句法选择策略上，甚至会影响我们对施动者和受动者的题元角色分配。模块化理论所倡导的



图 4　Knoeferle 等(2005)使用的视觉刺激示例。共包含三个角色，其中左侧为海盗，中间为拿着水桶正在清洗海盗的公主，右边为拿着画笔正在画公主的击剑者。

“封装”也在视觉情境范式的各类研究中受到了挑战，语言的加工并不是独立于其他信息的加工，而是与其他信息进行动态的即时交互。这些来自语言理解的研究都支持了基于制约的理论，语言的加工会实时的受到其他各类信息的影响和制约。

3.2 视觉信息对言语产生以及交流过程的影响

3.2.1 视觉信息对言语产生过程的影响

“听”别人说话并理解语言的过程会受到视觉信息的影响，同样，我们“说”的过程也同样会受到当前视觉画面或者场景的影响。言语产生过程多发生在某个特定视觉背景之下，个体需要对场景中的物体进行定位，与此同时也会提取物体的视觉特征以及相关的语言信息。有研究发现，个体在表达目标物体900 ms之前就会注视到相关的物体(Griffin & Bock, 2000)，视觉和语言信息的加工密不可分，需要跨通道合作才能完成整个言语产生过程。

已有研究发现，视觉刺激的不同特征会影响言语产生过程。例如，低水平的视觉特征会影响语言加工(Ostarek & Hüettig, 2017)。Rossion 和 Pourtois (2004)采用图片命名任务发现图片的颜色特征会影响图片的命名过程，带有颜色的图片要比黑白线条图片的识别速度快，命名也快，而且有颜色物体的命名一致性更高。Coco 和 Keller (2009)使用视觉情境范式，采用真实场景，通过改变所呈现场景的复杂程度和画面中人物数量，考察了视觉信息的复杂程度和特点对言语产生过程的影响。该研究发现视觉画面越复杂，人物越多，被试就会需要更多的时间来产生句子。阈下视觉刺激也会影响言语产生过程, Gleitman, January, Nappa 和 Trueswell (2007)发现在视觉画面呈现之前在目标位置呈现一个快速(60~75 ms)的注意捕捉信号(黑色方块)，结果发现，虽然被试报告并没有发现注意捕捉信号，但此位置出现的人物在句子产生过程中被作为主语的概率要更高。

Coco 和 Keller (2012)更为直接地观察到了视觉场景和言语产生之间的关系。以往多通道加工的研究发现，相对于两个不同的场景，两个相同的场景被试会有非常相似的扫视路径(scan pattern)。因此作者在其试验中，要求被试根据提示的线索(场景中所包含的物体)产生一个和场景相关的句子，考察对场景的扫视路径和句子产生之间协作的内在机制。结果显示，在言语产生的计划、编码以及产生阶段都发现场景的扫视路径相似和句子产生的相似有很高的相关，即对场景的扫视路径相同，产生的句子也会相似。Ferreira, Foucart 和 Engelhardt (2013)的实验 4 为了考察视觉情境范式中的预视阶段可以给被试提供何种信息，采用了言语产生范式，被试需要看视觉场景并在规定时间内猜测指导语内容。结果发现，在规定时间内，被试对指导语的猜测正确率显著大于了随机水平。这些研究都体现了视觉信息和语言产生过程之间的交互作用。

在言语产生领域中，另一个非常重要的问题是我们如何从图片中提取语义信息。这个问题同样也是人工智能领域的一个难题，如何让计算机“看图说话”？来自计算机领域的研究者 Vaidyanathan, Prud, Alm, Pelz 和 Haake (2015)以皮肤科专家为被试，采用经典的皮肤病图片作为实验材料试图建立语料库来让计算机“学会”提取图片中的语义信息。在其研究中，每个专家需要对 29 幅皮肤病的图像进行描述，并同时记录专家的眼动以及描述图片的声音数据。分析阶段，作者把眼动以及声音数据做成两个数据流并且严格地匹配形成一个“双维度语料库”，眼动数据作为视觉单元，录音数据作为语言单元，借助机器翻译的技术，成功地对图像进行了语义标注。经过训练的转换模块可以基于眼动数据(视觉信息)产生出对应的病情(语言单元)。表明视觉和语言之间存在语义上的联结，不同的视觉画面在注视期间会产生不同的包含语义信息的眼动数据，根据这些数据可以很好地预测出对应的语言单元。

3.2.2 视觉信息对言语交流的影响

言语交流作为复杂的语言加工现象，视觉信息的参与尤为重要。双方的视觉注意不仅会因为对方语言和视觉情境中物体的发生转移，还会影响对方的状态，继而影响语言加工过程。有研究发现，对话双方的口型、面部表情、反馈以及注视的变化等视觉信息都会影响双方的感知觉状态，与对话双方的语言加工产生交互作用，影响句法加工以及题元角色分配等过程(Carminati & Knoeferle, 2013; Garoufi, Staudte, Koller, & Crocker, 2016; Knoeferle & Kreysa, 2012; Kreysa, Knoeferle, & Nunneman, 2014)。例如，Carminati 和 Knoeferle (2013)的研究发现讲话者的带情绪的面部表情会影响听者的视觉注意以及语言理解过程。研究发

现对话双方对对方的视角进行捕捉可以帮助他们更好地理解言语的内容，并且对后续发言更好地计划(Tanenhaus & Brown-Schmidt, 2008)。Knoeferle 和 Kreysa (2012)发现被试能够根据讲话者注视的变化预测出讲话者将要提到的词汇。

对话双方所能共同获取的感知觉信息在言语交流中占有重要的地位。视觉呈现的物体可以同时呈现给对话双方，形成可以被双方同时观察到的视觉共享区域，研究发现个体会即时地将共享的视觉信息应用到目前的认知加工中。例如, Allopenna, Magnuson 和 Tanenhaus (1998)此前发现，被试在听到目标词“beaker (烧杯)”之后会看向和其起始音相同的竞争物“beetle (甲虫)”，被试听到的声音是通过耳机呈现的。有趣的是, Tanenhaus 和 Brown-Schmidt (2008)将实验过程变成听者和讲话者的交互对话过程，即讲话者直接对听者的对话，而非通过耳机呈现声音，并且双方可以同时看到一组物体。结果发现，视觉信息可以在对话中限制对话双方的知觉状态，会把语言的指涉范围(referential domain)限制在呈现的视觉物体上，语音竞争效应消失了，作者认为双方在对话过程中看到的视觉信息影响了双方的语言理解过程。Brown-Schmidt 和 Tanenhaus 等人针对听者与讲话者的视角做了大量研究，都表明对话双方的协作状态能够促进语言的加工过程，而这种协作状态大多情况下是有共同感知的视觉信息提供的。

视觉信息可以实时地为语言理解提供预测线索，提高交流效率。除去视觉场景来说，Huettig (2015)认为语言加工过程中预测性的存在的一个重要作用就是提高双方的交流效率。例如，在对话双方的交流过程中，经常补全对方语言的现象表明一方对另一方的言语产生内容进行了预测(Clark & Wilkesgibbs, 1986)。而视觉场景的呈现则更进一步提高了交流的效率，研究者采用视觉情境范式对语言加工中的预测性进行了考察，并发现在目标词还未出现之前就产生了指向目标物的眼动(Altmann & Kamide, 1999; Altmann, 2004; Altmann & Kamide, 2009; Hintz, Meyer, & Huettig, 2017; Trueswell & Thompson-Schill, 2016; Staub, Abbott, & Bogartz, 2012)。例如, Altmann 和 Kamide (1999)的研究让被试听句子“The boy will eat/move the cake (男孩将会吃掉/移动蛋糕)”的同时给被试呈现一个视觉场景，场景中包含“男孩”，目标物“蛋糕”以及其他干扰项。结果发现在被试听到“eat (吃)”的时候就已经有指向目标物“cake (蛋糕)”的眼跳出现，并且蛋糕得到了比其他干扰项更多的注视。对目标物的注视的原因不单是对句子中动词特征的分析，即“eat (吃)”后面要跟一个可食用的物体，而是源于视觉场景和语言表征的共同作用。视觉场景首先提供了线索，形成各种物体的视觉表征，这为后续句子中特定词汇的预测提供了基础，双方可根据场景中提供的视觉信息更好地对对方的讲话内容进行预测。

综上，言语产生过程中，个体从场景中提取语义信息，也因此言语产生过程会受到视觉画面的影响。不仅颜色、画面复杂程度等视觉信息会影响言语产生过程，讲话者的情绪面孔、注视变化等视觉信息都会影响听者感知觉状态，进而影响其语言加工过程。除此之外，视觉场景还能为对话双方的交流提供基础，对双方的言语产生内容进行预测，提高交流的效率。

3.2.3　视觉信息对语言理解过程的阻碍

以上研究表明，语言加工不是独立的单通道的过程，而是各种信息交互作用的结果。视觉、触觉、听觉等感觉通道都会影响语言的加工。视觉通道作为人类接受外界信息最主要的通道，会实时地帮助个体消除歧义句中的歧义，分配题元角色，对后续产生的词汇进行预测来协助语言的加工。但是这些影响并非都是促进作用。

首先，视觉信息的呈现会改变我们对语言本来的理解过程。例如, Pickering, Garrod 和 McElree (2004)等人指出视觉−情境范式中图片的呈现改变了语言理解过程。在他们给出的例子中，被试听指导语“In the morning Harry let out his dog Fido. In the evening he returned to find a starving beast (早上哈瑞放出了他那只叫费多的狗，傍晚回来的时候他发现了一只饥饿的野兽)”。对句子理解来说，“beast (野兽)”指向前半句提到的狗“Fido (费多)”，如果同时给被试呈现视觉图片“tiger (老虎)”，被试听到“beast”后可能会更多地注视“tiger”，即视觉信息的呈现会改变我们对语言的理解。除此之外，文字版的视觉情境范式中更能体现这种视觉信息的影响。在 Salverda 和 Tanenhaus (2010)的研究中，其在视觉画面中用文字代替物体给被试呈现目标词“bead”，竞争词“bear”以及无关项，同时听觉通道呈现目标词“bead”。结果发现被试

对竞争词的注视要显著多于无关项，体现出非常显著的竞争效应。Pickering 等人(2004)的研究同样也质疑这种语音竞争效应是由于视觉呈现的词汇影响了我们对目标词的识别，还是单纯语言加工里的竞争效应？

其次，视觉画面由于其较高的凸显性(salience)会影响个体语言表征的心理模拟过程。Altmann 和 Kamide (2009)的研究试图考察语言发生变化时，是否心理表征也会随之发生动态变化。被试听句子“The woman (will/is too lazy to) put the glass onto the table. Then, she will pick up the bottle, and pour the wine carefully into the glass (这个女人(会/太懒了以至于不会)把酒杯放到桌子上，然后，她会拿起酒瓶，把红酒倒入酒杯中)”的同时看一幅场景图片(如图 5)。结果发现，被试对“table (桌子)”的注视概率在移动酒杯的条件下要显著大于不移动酒杯的条件，表明语言引起了心理表征的模拟，在移动条件下，心理模拟的酒杯位置换到了桌子上。但在眼动数据上发现，不管是在哪个条件下，对于图片中酒杯的注视概率都要大于对于桌子的注视概率，这表明语言表征和视觉表征的两种表征机制产生了竞争效应，但由于视觉画面的突显性更大，所以会更占优势，比心理模拟指向的“table”得到更多的注视。在作者的实验 2 中，作者在听句子的过程中视觉画面在句子呈现的时候由灰屏替代，结果发现，对桌子的注视概率远远大于对酒杯的注视概率，语言表征的心理模拟过程则起了主导作用。



图 5　Altmann 和 Kamide (2009)使用的视觉刺激示例

最后，视觉画面的呈现会缩小个体对句子中特定词汇的指涉范围。在传统的研究语言加工的实验中，激活扩散模型认为听到某个词汇，心理词典中所有和此词汇相关的词汇都有可能得到激活。但如果有视觉场景或者图片呈现的话，被激活的词汇就会被限制在图片中所呈现的几个物体上。举例来说，Altmann 等人(1999)的研究中指导语为“the boy will eat the cake”，同时图片中只呈现一个可供食用的物体“cake (蛋糕)”，在听到词汇“eat”的时候，“cake”立即得到了注视。视觉信息限制了语言加工中可能被激活的条目，并不能反映整个心理词典的结构。

4　视觉信息在语言加工过程中所起的作用

以上概括了视觉信息影响语言加工多个方面的表现，不管是口语理解还是言语产生过程，都存在视觉信息与语言加工的跨通道交互作用。视觉信息为何会影响语言加工过程，在语言加工过程中起着何种作用，对这些问题探讨有助于揭示视觉信息与语言信息跨通道整合的机制。本节尝试对此进行一些综述和探讨。

首先，视觉信息会作为大脑的外部存储，有利于减少语言加工过程耗费的认知资源。Findlay 和 Gilchrist (2003)区分了两种视觉信息表征方式，被动视觉(passive vision)与主动视觉(active vision)。前者认为个体对视觉图像的理解过程是被动的，看过的图像作为视觉信息输入并存储在大脑中作为内部表征以供后续使用；后者则认为对视觉画面的理解是主动的，其加工的重要特点不是存储，而是个体对视觉画面外显的指向性注视，即，大脑会重新把注意转移到目标位置，以通过中央凹注视获取更精确的视觉信息。这两种观点最关键的区别是后续加工中视觉注意是否会转移。前者把视觉画面存储大脑中作为内部表征，后续提取的时候不需要重新注视，注意的转移是内隐的；后者并不存储视觉画面，只需存储位置信息，后续加工需要通过外显的注意转移来提取视觉信息。Findlay 等人认为视觉信息的加工模式是后者。Huettig, Gaskell 和 Quinlan (2004)同样认为后者符合认知系统的经济原则，这样视觉感知系统无需存储大量的视觉信息，而是把外部世界当成大脑的外部存储。这样来看，大脑中只需存储物体的空间位置信息并作为一个指针(pointer)，当语言加工需要提取相应视觉信息的时候，通过指

针指向特定位置来获取所需信息，大大减少了语言加工过程所耗费的认知资源。

使用视觉情境范式的研究同样证实了这个观点。在视觉情境范式中听觉刺激呈现之前会有对视觉画面的预视阶段，结果也都发现当听到目标词的时候会出现指向目标物的注意转移。如果视觉画面存储在大脑中，则无需眼动便可获取信息，这种注意的转移恰恰表明大脑把外部世界当成了外部存储。Altmann (2004)改变了视觉情境范式，画面在预视阶段之后消失并同时呈现空白屏幕和听觉刺激，作者称之为“空屏范式(the blank screen paradigm)”。结果发现被试依然会看向目标物曾经出现过的位置，这个结果和上面的假设是吻合的，被试储存了物体的空间位置信息，并将视觉画面作为了大脑的外部存储。不仅如此，更有研究发现和目标词语义相关的物体在空屏范式下也会引起指向目标物的眼动(De Groot, Huettig, & Olivers, 2016)。Richardson 和 Spivey (2000)认为，这种空间位置信息的存储是视觉系统利用眼动协调(Oculomotor coordinate)来实现的，视觉系统并非直接记录整个场景，而是指引眼睛移动到相应的坐标提取相应的场景。视觉信息和语言的加工过程可能存在这样一个系统：语言会指向相应的位置，并且只有当眼睛到达目标位置的时候关于这个位置的具体信息才会提取出来。

其次，正如语言信息可以影响我们对物体的分类一样，在语言习得过程中，视觉信息同样可以塑造语言的加工过程。目前很多研究者强调了婴幼儿与成人中语言加工的多通道特性(Mani & Schneider, 2013; Yeung & Nazzi, 2014; Yeung & Werker, 2009)。语言加工属高级认知过程，与之相比，视觉信息在婴幼儿的早期发展中占有更为重要的地位。有关儿童词汇识别的研究发现，幼儿听到一个词汇的时候可以提取与这个词汇相联系物体的感知觉信息(Arias-Trejo & Plunkett, 2009; Johnson & Huettig, 2011; Johnson, McQueen, & Huettig, 2011; Mani, Johnson, McQueen, & Huettig, 2013)。更有研究发现，儿童在目标词出现之前就会激活其形状信息，表现在对和目标词形状相似物体更多的注视上(Bobb, Huettig, & Mani, 2016)。Yeung 和 Werker (2009)发现仅仅是教婴儿区分两种形状不同的物体和两种声音之间的联系就可以帮助婴儿更好地区分开两种声音。这些研究都表明视觉信息在语言习得和加工中起着重要的作用，感知觉信息与听觉语言信息的共同激活可以帮助儿童在听到一个词汇的时候，在其所处的场景中更快地寻找到匹配的物体。很多研究采用视觉情境范式对儿童语言理解发展进行了多方面的考察，发现虽然儿童可以利用视觉信息来帮助区分声音或是协助语言的习得，但儿童在视觉和语言的整合功能上和成人依旧存在差异(Bunger, Skordos, Trueswell, & Papafragou, 2016; Huang & Snedeker, 2009, 2011; Melissa, Snedeker, & Schulz, 2017)。例如，有研究发现儿童在概念表征和句法解歧上与成人也存在着显著的差异(Pluciennicka, Coello, & Kalénine, 2016)。也有研究发现在第二语言的习得上，二语者和母语者表现为不同的影响模式(Ito, Pickering, & Corley, 2018; Noh & Lee, 2017; Pozzan & Trueswell, 2016)。

最后，视觉信息可以移除或者降低句子加工中的加工困难。在本文第二部分已列出多种视觉信息可以移除或者降低歧义句中的加工困难的例证。例如，Tanenhaus 等人(1995)使用视觉背景消解了句子的暂时歧义，在双表征物语境下，被试的错误注视概率减少。笔者认为，惊异理论(Surprisal theory)可以很好地解释视觉信息对语言加工中句法歧义消解的影响，并且已有研究使用惊异理论解释句子加工中的句法选择策略(Staub & Clifton, 2006)。惊异理论是计算语言学家 Hale (2001)提出的一个概念，用来描述句子理解过程中遇到某个词后产生的加工困难或者说认知负担，惊异系数(surprisal)的高低决定了句法加工难度。举例来说，Tanenhaus 等人的研究使用的局部歧义句“put the apple on the towel in the box”由于“on the towel”的歧义现象，在遇到“in”的时候会产生加工困难。双表征物语境的视觉画面中，错误注视概率减少，表明视觉背景的呈现减少了介词短语“in the box”的惊异系数。视觉背景能够影响句法加工策略的过程可以被看成降低句子加工困难的过程。可惜的是，由于视觉背景难以量化，计算句子中惊异系数的改变也是一个非常大的难题，此类研究少之又少，笔者只发现有研究考察了世界知识对于句子中惊异系数的影响(Venhuizen, Brouwer, & Crocker, 2016)。

总之，视觉信息在语言的加工过程中扮演着非常重要的角色，这不仅表现在成人身上，在儿

童语言的不同发展阶段中也占有重要位置。视觉信息以其非常高的突显性不仅能够为其他认知过程提供大量的信息作为加工的基础，还可以实时地参与到认知加工中来。对于语言加工来说：首先，视觉信息可以作为大脑的外部存储降低语言加工的认知负担，听到相应词汇再去相应位置获取更精确的信息，增加了视觉与语言加工之间的互动；其次，视觉信息极大地促进了儿童的语言习得过程，帮助儿童更好地把语言以及生活中的物体进行匹配；再次，视觉信息可以帮助我们降低句子加工中遇到的加工困难；最后，如第 3 部分提到的视觉信息对后续词汇的预测以及对言语交流的影响上看，视觉信息还可以帮助我们更好地预测出句子后续将会输入的词汇，使对话双方更好地进行交流，当双方处在同一个场景中的时候，能够减少语言产生的负担。这些都是视觉信息和语言加工整合的原因，但目前还没有对为什么视觉信息和语言加工之间的交互有确切的解释，本节希望能够为揭示这种跨通道整合的内在机制提供一些思路。

5 视觉影响语言加工的计算机模型

理论假设和实验论证是实践应用的基础，科技的发展使人工智能技术越来越多地出现在生活的方方面面，如何将理论基础运用在科技实践上是目前我们面临的重要问题。很多研究者开始尝试着通过构建计算机模型来模拟跨通道的交互作用，以促进人工智能领域的发展。因此，对目前“视觉信息影响语言加工计算模型”的梳理不仅有利于我们对这种交互机制的全面理解，而且有利于提起对实践应用的重视。目前的模型大多是对词汇水平的视觉−情境范式的模拟，旨在揭示在口语理解过程中语义、语音、字形以及视觉特征等激活的时间进程。这类研究的计算机模拟相对较为成熟，不少研究使用之前的计算模型对视觉−情境范式的研究进行模拟取得了比较可靠的数据(McClelland, Mirman, Bolger, & Khaitan, 2014; Smith, Monaghan & Huettig, 2013; Smith, Monaghan & Huettig, 2014, 2017)。目前用来模拟词汇水平的视觉−情境范式比较成熟的模型有“工作记忆模型(working memory model)” (Huettig et al., 2011), “中心辐射模型(Hub and Spoke model, H&S)” (Dilkina, McClelland & Plaut, 2010; Smith et al., 2013)以及一种神经网络模型“简单递归网络模型(simple recurrent network model, SRN)” (Elman, 1990)。

但以上的这些模型重点都放在语言如何引导视觉注意上，并未揭示视觉信息如何影响语言加工过程。据笔者了解，心理语言学中还没有研究者对视觉背景影响句法歧义消解的过程建立模型。目前一些来自计算语言学的研究者尝试把视觉信息与听觉语言信息在语义层面建立接口来模拟这类实验，但往往停留在一个描述性的层面(Baumgärtner, Beuck, & Menzel, 2012; McCrae, 2009)。例如，Venhuizen 等(2016)通过建立向量模型，考察了世界知识(world knowledge)对于语言理解中加工困难的影响，但是其分析也主要是把世界知识转换成事件发生的先后顺序，来估测不同的先后顺序下句子中每个词汇出现的概率，并没有提出一个系统的模型对整个过程进行模拟。这类模型的困难点主要是难以将“情境”这种高水平的信息量化，使得计算机模拟相对较为困难。这部分主要简单介绍一种关于视觉信息如何影响语言加工的模型。

来自视觉情境范式的研究认为这种视觉和语言跨通道的交互作用发生在语义表征层面(Altmann, 2004)。这和 Jackendoff 的概念语义学理论相符合，Jackendoff 根据一系列的语言和认知科学的证据认为，存在一个心理表征层面“概念结构”，这个层面是各种认知信息交流的接口，语言、感知觉、运动等信息都会在这个层面发生交互作用(Jackendoff, 1983)。McCrae (2009)基于此假设构建了一个模型，来试图模拟视觉背景信息与句法解析之间的交互作用。

McCrae 模型最终目的是建立加入视觉背景信息后的句法解析器(parser)，即在有视觉信息的影响下进行句法解析。模型的建立首先需要一种句法解析器对句子的句法进行解析；其次，在此解析器上建立接口以输入视觉信息。作者和其合作者之前构建了一个权重制约的依存句法解析器(WCDG, Weighted constraint dependency parser)，此解析器提供了包含多种非语言信息的一般性接口，对于研究视觉信息对句法加工的影响有很大优势。作者借助 WCDG，以 Jackendoff 的理论作为基础，构建了此模型来模拟视觉信息与句法解析之间的交互作用。模型由三个模块构成，分别为：语言模块、概念结构模块以及视觉感知模块(如图 6)。

Syntactic representation (句法表征)↔ Semantic representation (语义表征)
↑
Visual Scene Context (视觉背景)

图 6　模型组成以及模块间的交互作用(McCrae, 2009)。其中句法表征为语言模块，语义表征为概念结构模块，视觉背景为视觉感知模块。

由于视觉信息的难以量化，该模型对视觉信息模块的处理是把视觉信息描述的事件简化成题元角色的分配，即施动者和受动者的角色分配。语言信息模块中，作者采用德语中的歧义句式“谁对谁做了什么(who did what to whom)”，最终通过 WCDG 来接入视觉模块的输入的角色分配的信息流，并和语言模块中的角色分配信息进行匹配，最后输出模型的句法解析结果。作者在实际的模拟中，单独采用语言模块和采用加入视觉背景信息的模块对歧义句进行解析的结果是不同的，视觉信息的加入可以改变本身的句法选择策略，成功地对此类现象进行了模拟。

6　总结与展望

本综述重点梳理了视觉信息如何影响语言加工过程的研究，从口语理解、言语产生以及言语交流等方面概括了视觉信息影响语言加工过程的表现。总体来看，语言加工的过程并不是独立进行的，模块化理论中“封装”起来的句法加工模块也受到了挑战。采用视觉−情境范式考察口语理解，言语产生的研究都发现视觉场景、动作特点、情节记忆以及事件等信息可以即时的影响语言加工过程，语言的加工是汇集了各类不同通道的信息实时交互作用的结果。视觉场景不仅可以作为大脑的外部存储器降低我们语言加工过程中的认知资源消耗，而且可以促进语言习得过程，降低我们语言加工过程中遇到的加工困难，提高语言加工效率，促进言语交流过程。

不仅很多研究者对视觉信息影响语言加工的现象展开了研究，而且也同样有很多研究考察了语言加工对视觉注意的引导过程，视觉与语言之间的交互机制的研究和解决是揭示人类跨通道整合机制的关键环节。但目前这个领域还有很多亟待解决的问题，将来的研究应该围绕揭示视觉与语言整合的内在机制，如何利用现有研究理论来指导儿童语言发展中的视听整合过程，以及如何促进人工智能的发展这三个大问题来展开，解决这些问题将会极大促进我们对人类认知过程的全面理解。

第一，揭示视觉与语言整合的内在机制。已有研究开始关注不同通道信息交互作用的神经机制。例如，Hagoort (2005)构建了一个语言加工的神经结构模型，主要从布洛卡区着手分别阐述了语音、句法、语义在神经结构上的整合过程，并且强调了左侧额下回(LIFG, left inferior frontal gyrus)对非语言信息(例如手势)和语言信息的整合的重要作用。Peeters, Snijders, Hagoort 和 Özyürek (2017)利用事件相关功能磁共振成像技术也同样发现左侧额下回与双侧颞中回在口语和视觉情境交互中的重要作用。但是，这些研究只考察了“听声识物”过程中的神经机制，视觉信息与句法加工，语义加工等交互作用的神经机制将会是未来非常重要的研究课题。

第二，如何利用现有研究结果来指导儿童语言发展中的视听整合过程。多个研究发现，婴幼儿语言加工同样有多通道特性，视觉信息可以塑造语言的加工过程。但儿童在视觉和语言的整合功能上和成人依旧存在差异，并不能很好地利用视觉信息来进行语言加工，因此，如何使用现有的研究结果与理论训练与干预儿童的语言习得过程，以提高其语言加工效率，促进儿童认知发展显得尤为重要。

第三，视觉通道是多种感觉通道中的一种，是人类获取信息最主要的通道。而对于盲人或者有视觉缺陷的人，听觉和触觉则是最直接和有效的。因此对于视觉和语言加工交互机制的揭示有助于推进其他感觉通道和语言加工交互作用的研究。例如，在一些情景下，视觉信息起到的是语境的作用，若以其他通道呈现，也可能同样会影响到语言的加工过程，起到相同的效果。为了提高这类人群的生活质量，这类研究必将有广阔的发展前景。

第四，人工智能领域在现代信息技术的带领下飞速发展，已经慢慢进入到现代生活当中，在

各个行业也得到了广泛应用。人工智能技术通过多通道的整合技术能够实现更多更全面的功能。然而，关于视觉信息与语言加工交互作用的模型依旧还是短板，如何量化视觉信息，并快速地和语言进行匹配，目前的计算模型都尚未解决这些问题。因此视觉和语言交互作用机制的揭示可以使我们了解这些信息如何共同作用实现视听整合，从而为人工智能的进一步发展提供科学依据。

参考文献

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(38), 419–439.

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: The "blank screen paradigm." *Cognition*, *93*(2), 79–87.

Altmann, G. T. M., Garnham, A., & Dennis, Y. (1992). Avoiding the garden path: Eye movements in context. *Journal of Memory and Language*, *31*(5), 685–712.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Altmann, G. T. M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, *111*(1), 55–71.

Arias-Trejo, N., & Plunkett, K. (2009). Lexical-semantic priming effects during infancy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3633–3647.

Baumgärtner, C., Beuck, N., & Menzel, W. (2012). An architecture for incremental information fusion of cross-modal representations. *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 498–503.

Beauchamp, M. S. (2016). Chapter 42-Audiovisual speech integration: Neural substrates and behavior. *Neurobiology of Language*, (2011), 515–526.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *The Journal of Neuroscience*, *17*(1), 353–362.

Bobb, S. C., Huettig, F., & Mani, N. (2016). Predicting visual information during sentence processing: Toddlers activate an object's shape before it is mentioned. *Journal of Experimental Child Psychology*, *151*, 51–64.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science*, 32(4), 643−684.

Bunger, A., Skordos, D., Trueswell, J. C., & Papafragou, A. (2016). How children and adults encode causative events cross-linguistically: Implications for language production and attention. *Language, Cognition and Neuroscience*, 31(8), 1015−1037.

Carminati, M. N., & Knoeferle, P. (2013). Effects of speaker emotional facial expression and listener age on incremental sentence processing. *PLoS ONE*, *8*(9), e72559.

Chambers, C. G., & Juan, V. S. (2008). Perception and presupposition in real-time language comprehension: Insights from anticipatory processing. *Cognition*, *108*(1), 26–50.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Carlson, G. N., & Filip, H. (1998). Words and worlds: The construction of context for definite reference. In *Proceedings of the 20th Annual Meeting of the Cognitive Science Society, Mahwah, NJ: Lawrence Erlbaum* (pp. 220−225).

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(3), 687–696.

Chen, P-H., & Tsai, J-L. (2015). The influence of syntactic category and semantic constraints on lexical ambiguity resolution: An eye movement study of processing Chinese homographs. *Language and Linguistics*, 16(4), 555−586.

Clark, H. H., & Wilkes-gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Coco, M. I., & Keller, F. (2009). The impact of visual information on reference assignment in sentence production. *Conference of the Cognitive Science Society,* 274–279.

Coco, M. I., & Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, *36*(7), 1204−1223.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107.

De Groot, F., Huettig, F., & Olivers, C. N. L. (2016). Revisiting the looking at nothing phenomenon: Visual and semantic biases in memory search. *Visual Cognition*, 24, 226–245.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2010). Are there mental lexicons? The role of semantics in lexical decision. *Brain Research*, *1365*, 66–81.

Elman, J. L. (1990). Finding structure in time. *Cognitive*

Science, *14*(2), 179–211.

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language*, 69(3), 165–182.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. US: Oxford University Press.

Fodor, J. A. (1983). *The modularity of mind*. MIT press Cambridge.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2), 178−210.

Garoufi, K., Staudte, M., Koller, A., & Crocker, M. W. (2016). Exploiting listener gaze to improve situated communication in dynamic virtual environments. *Cognitive Science*, 40(7), 1671–1703.

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*(4), 544–569.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*(4), 274−279.

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual. Review. Neuroscience*, *27*, 649−677.

Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Extraction of event roles from visual scenes is rapid, automatic, and interacts with higher-level visual processing. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (Vol. 73).

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences, 9(9)*, 416–423.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1−8).

Heinrich, S., & Wermter, S. (2018). Interactive natural language acquisition in a multi-modal recurrent neural architecture. *Connection Science, 30*(1), 99-133.

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1352−1374. doi:10.1037/xlm0000388.

Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, *45*(6), 1723–1739.

Huang, Y. T., & Snedeker, J. (2011). Logic and conversation revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes*, *26*(8), 1161–1172.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135.

Huettig, F., Gaskell, M. G., & Quinlan, P. T. (2004). How speech processing affects our attention to visually similar objects: Shape competitor effects and the visual world paradigm. In *Proceedings of the 26th annual meeting of the Cognitive Science Society* (pp. 607–612).

Huettig, F., Olivers, C. N. L., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta psychologica*, *137*(2), 138−150.

Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, *98*, 1−11.

Jackendoff, R. (1983). *Semantics and cognition* (Vol. 8). MIT press.

Johnson, E. K., & Huettig, F. (2011). Eye movements during language-mediated visual search reveal a strong link between overt visual attention and lexical processing in 36-month-olds. *Psychological Research*, *75*(1), 35–42.

Johnson, E. K., McQueen, J. M., & Huettig, F. (2011). Toddlers' language-mediated visual search: They need not have the words for it. *The Quarterly Journal of Experimental Psychology*, *64*(9), 1672–1682.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, *95*(1), 95–127.

Knoeferle, P., & Guerra, E. (2016). Visually situated language comprehension. *Language & Linguistics Compass*, *10*(2), 66–82.

Knoeferle, P., & Kreysa, H. (2012). Can speaker gaze modulate syntactic structuring and thematic role assignment during spoken sentence comprehension? *Frontiers in Psychology*, *3*, 538.

Kreysa, H., Knoeferle, P., & Nunneman, E. M. (2014). Effects of speaker gaze versus depicted actions on visual attention during sentence comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Kuchenbuch, A., Paraskevopoulos, E., Herholz, S. C., & Pantev, C. (2014). Audio-tactile integration and the influence of musical training. *PloS One*, *9*(1), e85743.

Lee, R., Chambers, C. G., Huettig, F., & Ganea, P. A. (2017). Children's semantic and world knowledge overrides

fictional information during anticipatory linguistic processing. In *The 39th Annual Meeting of the Cognitive Science Society* (CogSci 2017) (pp. 730−735).

Leonard, M. K., & Chang, E. F. (2014). Dynamic speech representations in the human temporal lobe. *Trends in Cognitive Sciences*, 18(9), 472–479.

Linzen, T., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 40(6), 1382−1411.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, *32*(5), 692–715.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676–703.

Mani, N., Johnson, E., McQueen, J. M., & Huettig, F. (2013). How yellow is your banana? Toddlers' language-mediated visual search in referent-present tasks. *Developmental Psychology*, *49*(6), 1036–1044.

Mani, N., & Schneider, S. (2013). Speaker identity supports phonetic category learning. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 623–629.

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science*, *189*(4198), 226–228.

McClelland, J. L., Mirman, D., Bolger, D. J., & Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive Science*, *38*(6), 1139–1189.

McCrae, P. (2009). A model for the cross-modal influence of visual context upon language processing. *International Conference Recent Advances in Natural Language Processing (RANLP 09, Borovets, Bulgaria)*, 230–235.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Melissa, K., Snedeker, J., & Schulz, L. (2017). Linking language and events: Spatiotemporal cues drive children's expectations about the meanings of novel transitive verbs. *Language Learning and Development*, *13*(1), 1–23.

Milburn, E., Warren, T., & Dickey, M. W. (2015). World knowledge affects prediction as quickly as selectional restrictions: Evidence from the visual world paradigm. *Language, Cognition and Neuroscience*, *31*(4), 536–548.

Ng, H. G., Anton, P., Brügger, M., Churamani, N., Fließwasser, E., Hummel, T., ... Wermter, S. (2017). Hey Robot, Why don't you talk to me. *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, Lisbon, Portugal.*

Noh, Y., & Lee, M. (2017). The impact of inhibitory controls on anticipatory sentence processing in L2. *Journal of Cognitive Science*, *18*(1), 21–41.

Nozari, N., Trueswell, J. C., & Thompson-Schill, S. L. (2016). The interplay of local attraction, context and domain-general cognitive control in activation and suppression of semantic distractors during sentence comprehension. *Psychonomic Bulletin & Review*, *23*(6), 1942–1953.

Ostarek, M., & Hüettig, F. (2017). Spoken words can make the invisible visible – Testing the involvement of low-level visual representations in spoken word processing. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(3), 499−508.

Peeters, D., Snijders, T. M., Hagoort, P., & Özyürek, A. (2017). Linking language to the visual world: Neural correlates of comprehending verbal reference to objects through pointing and visual cues. *Neuropsychologia*, *95*, 21–29.

Pickering, M. J., Garrod, S., & McElree, B. (2004). Interactions of language and vision restrict "visual world" interpretations.

Pluciennicka, E., Coello, Y., & Kalénine, S. (2016). Development of implicit processing of thematic and functional similarity relations during manipulable artifact object identification: Evidence from eye-tracking in the Visual World Paradigm. *Cognitive Development*, *38*, 75–88.

Pozzan, L., & Trueswell, J. C. (2016). Second language processing and revision of garden-path sentences: A visual word study. *Bilingualism: Language and Cognition*, *19*(3), 636−643.

Eggermont, J. J. (2017). *Hearing loss: Causes, prevention, and treatment.* Academic Press.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition*, *76*(3), 269–295.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217–236.

Salverda, A. P., & Tanenhaus, M. K. (2010). Tracking the time course of orthographic information in spoken-word recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 36*(5), 1108–1117.

Smith, A. C., Monaghan, P., & Huettig, F. (2017). The multimodal nature of spoken word processing in the visual world: Testing the predictions of alternative models of multimodal integration. *Journal of Memory and Language*, *93*, 276–303.

Smith, A. C., Monaghan, P., & Huettig, F. (2014). Modelling

language–Vision interactions in the hub and spoke framework. *Computational Models of Cognitive Processes*, 3−16.

Smith, A. C., Monaghan, P., & Huettig, F. (2014). A comprehensive model of spoken word recognition must be multimodal: Evidence from studies of language-mediated visual attention. In *36th Annual Meeting of the Cognitive Science Society (CogSci 2014)*. Cognitive Science Society.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, *49*(3), 238–299.

Staub, A., Abbott, M., & Bogartz, R. S. (2012). Linguistically guided anticipatory eye movements in scene viewing. *Visual Cognition*, *20*(8), 922–946.

Staub, A., & Clifton Jr, C. (2006). Syntactic prediction in language comprehension: Evidence from either...or. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(2), 425–436.

Tanenhaus, M. K., & Brown-Schmidt, S. (2008). Language processing in the natural world. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 1105−1122.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Vaidyanathan, P., Prud'hommeaux, E., Alm, C. O., Pelz, J. B., & Haake, A. R. (2015). Alignment of eye movements and spoken language for semantic image understanding. *Proceedings of the 11th International Conference on Computational Semantics*, 76–81.

van Bergen, G., & Flecken, M. (2017). Putting things in new places: Linguistic experience modulates the predictive power of placement verb semantics. *Journal of Memory and Language*, *92*, 26–42.

Venhuizen, N. J., Brouwer, H., & Crocker, M. (2016). When the food arrives before the menu: Modeling event-driven surprisal in language comprehension. In *Abstract Presented at Events in Language and Cognition, Pre-CUNY Workshop on Event Structure (Gainesville, FL).*

Yeung, H. H., & Nazzi, T. (2014). Object labeling influences infant phonetic learning and generalization. *Cognition*, *132*(2), 151–163.

Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, *113*(2), 234–243.

Cross-modal integration of audiovisual information in language processing

HAN Haibin; XU Pingping; QU Qingqing; CHENG Xi; LI Xingshan

(*Institute of Psychology, Chinese Academy of Sciences, Beijing* 100101, *China*)

(*University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: In daily life, the use of language often occurs in a visual context. A large number of cognitive science studies have shown that visual and linguistic information processing modules do not work independently, but have complex interactions. The present paper centers on the impact of visual information on language processing, and first reviews research progress on the impact of visual information on speech comprehension, speech production and verbal communication. Secondly, the mechanism of visual information affecting language processing is discussed. Finally, computational models of visually situated language processing are reviewed, and the future research directions are prospected.

Key words: visual information processing; language processing; speech comprehension; speech production; verbal communication